Statistical Learning Theory (2)

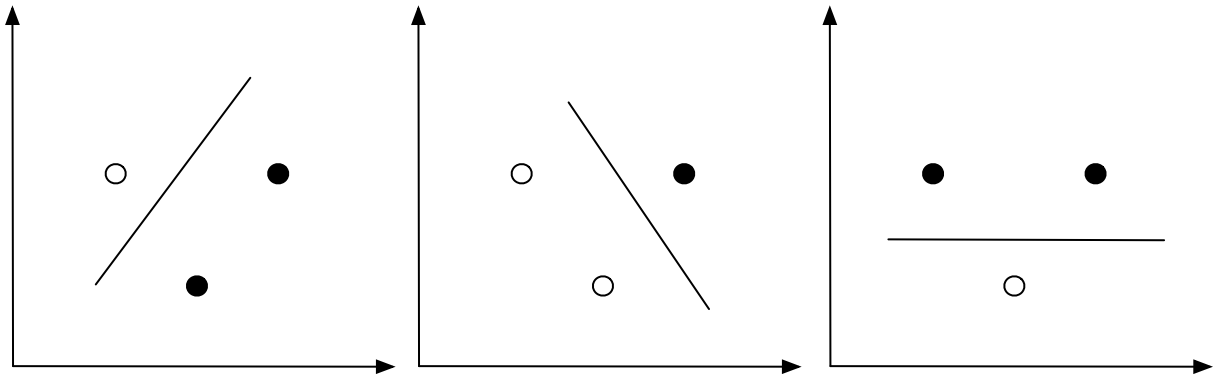
Manel Martínez-Ramón

ECE, UNM

October, 2018

The Vapnik-Chervonenkis dimension

- Consider the space $\mathbb{R}^2$ and a set of vectors. In this space, only three points can be *linearly* shattered in all possible ways if two are linearly independent.



If a fourth point is added, then not all possibilities can be achieved with a linear function because in any case only two will be linearly independent.

Theorem

- Consider some set of m points in $\mathbb{R}^n$. Choose any one of the points as origin. Then the m points can be shattered by oriented hyperplanes if and only if the position vectors of the remaining points are linearly independent.

Corollary

- The maximum number of vectors that can be shattered by a hyperplane in a space $\mathbb{R}^n$ is $n+1$, since we can always choose $n+1$ points, and then choose one of the points as origin, such that the position vectors of the remaining n points are linearly independent, but can never choose $n+2$ such points since no $n+1$ vectors in $\mathbb{R}^n$ can be linearly independent.

(See the paper by Burges for a proof.)

The maximum number of vectors that can be shattered by a hyperplane is called the **VC dimension**.

- The VC dimension of a hyperplane in a space of dimension n is $h = n + 1$
- The VC dimension gives us a measure of the complexity of linear functions.
- If the VC dimension of an estimator is higher than the number of vectors to be classified, then the estimator is guaranteed to overfit if an empirical risk is minimized over the data, since all vectors will be correctly classified regardless of their statistical properties.

Theorem (Vapnik, 1995)

- Define the linear empirical risk as

$$R_{emp}(\boldsymbol{\alpha}) = \frac{1}{2N} \sum_{n=1}^{N} |y - f(\mathbf{x}, \boldsymbol{\alpha})|$$

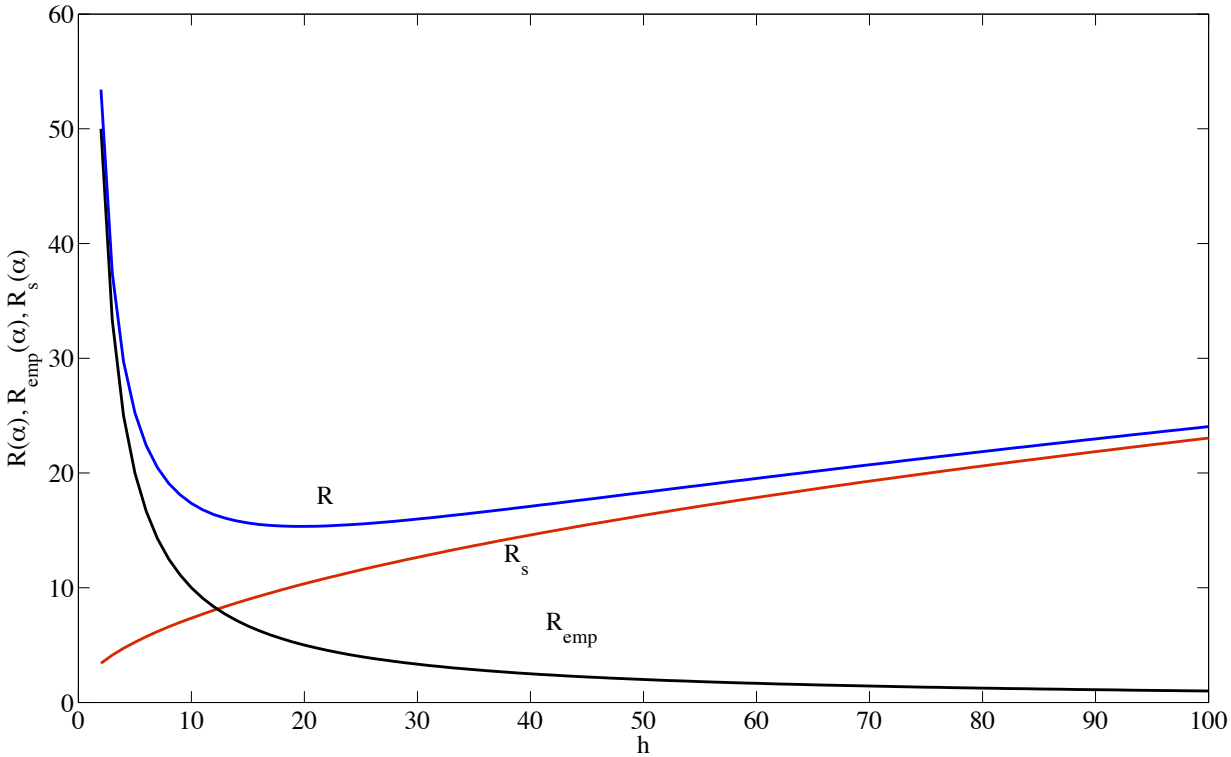
  where $f(\cdot)$ is defined so that the loss function $|y - f(\mathbf{x}, \boldsymbol{\alpha})|$ can only take the values 0 or 1.

- Then, with probability $1 - \eta$ the following bound holds

$$R(\boldsymbol{\alpha}) \leq R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h(log(2N/h) + 1) - log(\eta/4)}{N}}$$

- The second term of the right side is called the Structural risk (R_s).

Graphically, for $N = 100$, $\eta = 10^{-3}$



We will find a real estimation of this graph in the homework of this chapter.

Remarks

$$R(\boldsymbol{\alpha}) \leq R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h(log(2N/h)+1) - log(\eta/4)}{N}}$$

- This is a bound on the risk with probability $1-\eta$, so it is not guaranteed.
- The bound is independent of the probability distribution.
- While the left side is not computable, the right one can be easily computed provided the knowledge of h.

The inductive *Principle of Structural Risk Minimization* consists then on chosing a machine whose dimension h is sufficiently small, so the bound on the risk is minimized.

- As an additional remark, Vapnik's theorem is not restricted to any particular class of machines, though we just defined h for linear ones.
- Nevertheless, linear machines are of particular interest because its VC dimension can be computed and minimized.
- Nonlinear extensions of linear machines can be constructed with the Kernel trick to be presented in next chapter.
- The Kernel trick preserves the linearity of machines through a nonlinear transformation to higher dimension Hilbert spaces.

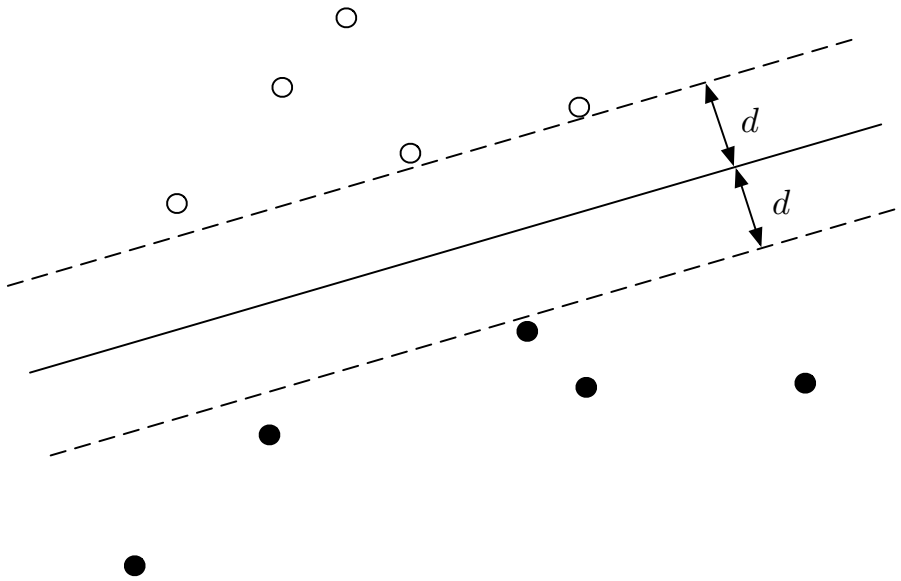
Minimization of the VC dimension of a linear machine (Vapnik, 1995)

- The VC dimension of a separating hyperplane is minimized if the norm of its parameters is minimized.
- Consider the estimation function

$$f(x) = \text{sign}(\mathbf{w}^\top x + b)$$

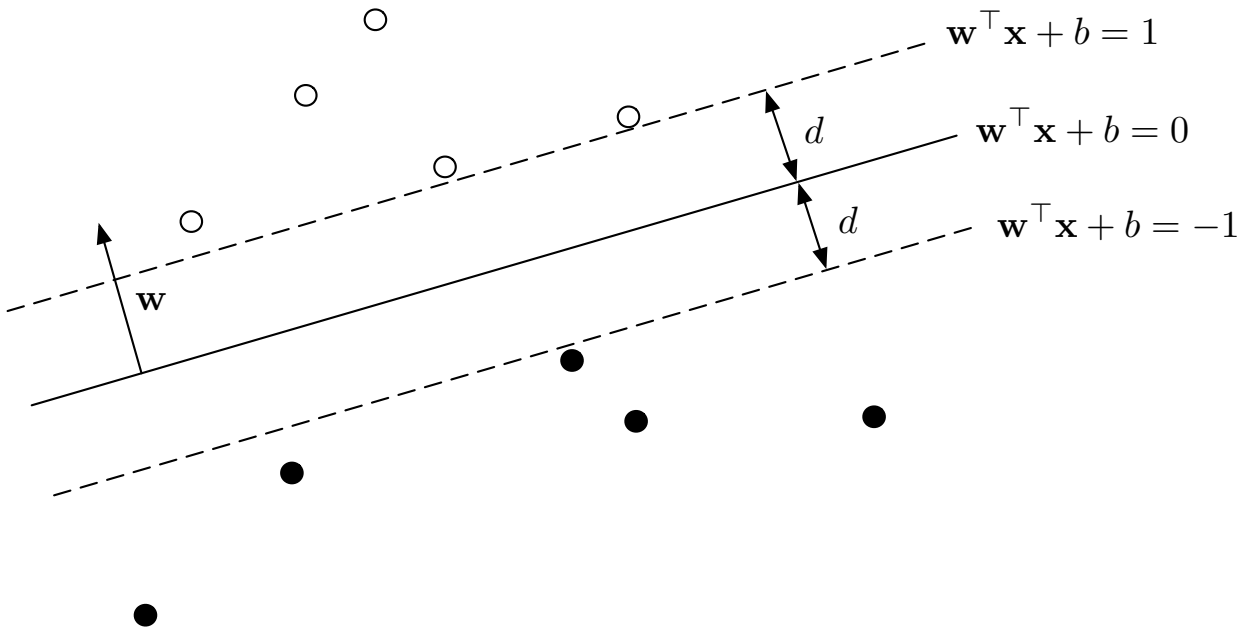
- and a set of data $\mathcal{D} : \{\mathbf{x}_1, y_1, \cdots \mathbf{x}_N, y_N\}$ that can be linearly classified according to their labels $y_n \in \{-1, 1\}$.

- A way to limit the VC dimension so the plane can only separate the patterns $\mathbf{x}_n$ according to their labels is to *maximize the margin between the two classes.*



In other words, we will place the plane so we maximize its distance to the closest patterns, while correctly classifying them.

- Now we define the margins as functions $\mathbf{w}^\top \mathbf{x} + b = \pm 1$ and its distance to the separating hyperplane as d.
- Note also that $\mathbf{w}$ is a vector normal to the hyperplane.

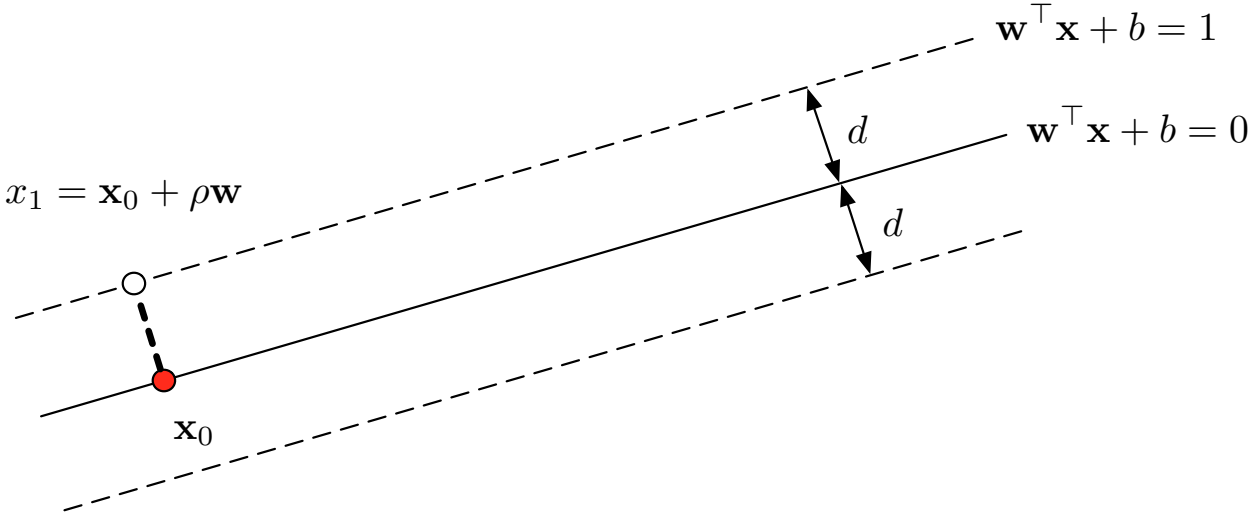


- In order to compute d we define a point $\mathbf{x}_0$ such that

$$\mathbf{w}^\top \mathbf{x}_0 + b = 0$$

From this point we trace a line perpendicular to the plane

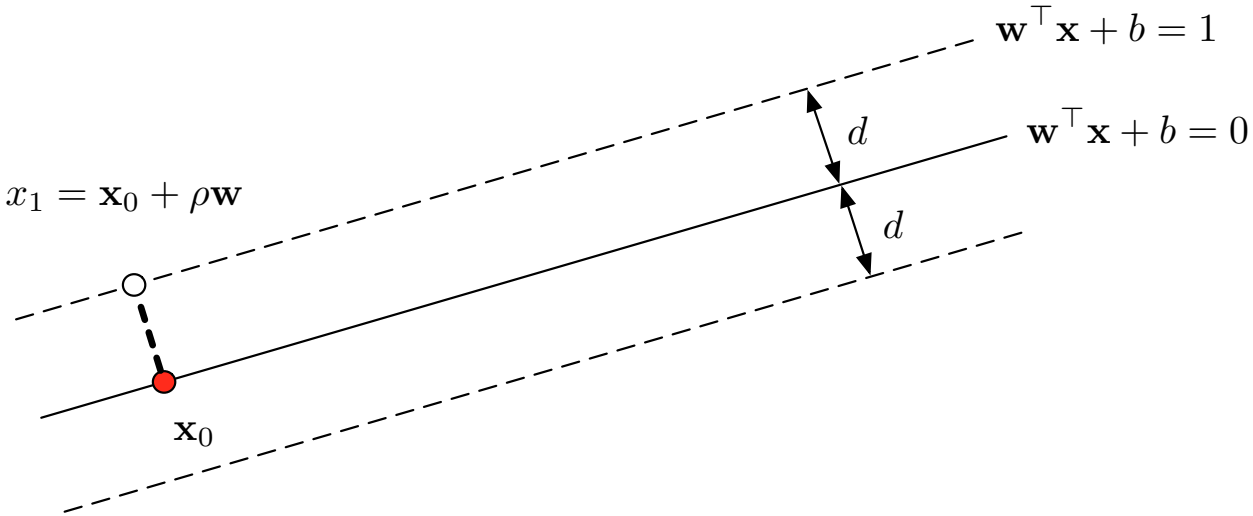
$$\mathbf{x} = \mathbf{x}_0 + \rho \mathbf{w}$$



- Its intersection with the upper margin is a point $\mathbf{x}_1$ such that

$$\mathbf{w}^\top \mathbf{x}_1 + b = 1$$

$$\mathbf{x}_1 = \mathbf{x}_0 + \rho \mathbf{w}$$



- The distance is then $d = ||\mathbf{x}_1 - \mathbf{x}_0|| = \rho ||\mathbf{w}||$.

- Now using equations

$$\mathbf{w}^\top \mathbf{x}_1 + b = 1$$
$$\mathbf{x}_1 = \mathbf{x}_0 + \rho \mathbf{w}$$

we straightforwardly find

$$\mathbf{w}^\top (\mathbf{x}_0 + \rho \mathbf{w}) + b = 1$$
$$\mathbf{w}^\top \mathbf{x}_0 + b + \rho ||\mathbf{w}||^2 = 1$$

Since we defined $\mathbf{x}_0$ such that $\mathbf{w}^\top \mathbf{x}_0 + b = 0$

$$\rho = \frac{1}{||\mathbf{w}||^2}$$

and

$$d = \rho ||\mathbf{w}|| = \frac{1}{||\mathbf{w}||}$$

- The margin $d = \frac{1}{||\mathbf{w}||}$ says that maximizing the margin is equivalent to minimizing the norm of the parameters subject to the constraint of correctly classifying all samples, this is

$$\begin{aligned} &\text{minimize } ||\mathbf{w}||^2 \\ &\text{subject to } y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) > 1 \end{aligned}$$

- This, in turn, constructs the machine with the minimum possible complexity by restricting the possible shatters of the plane to 1.

This is *almost* what we call the Support Vector Machine, **but not yet**.

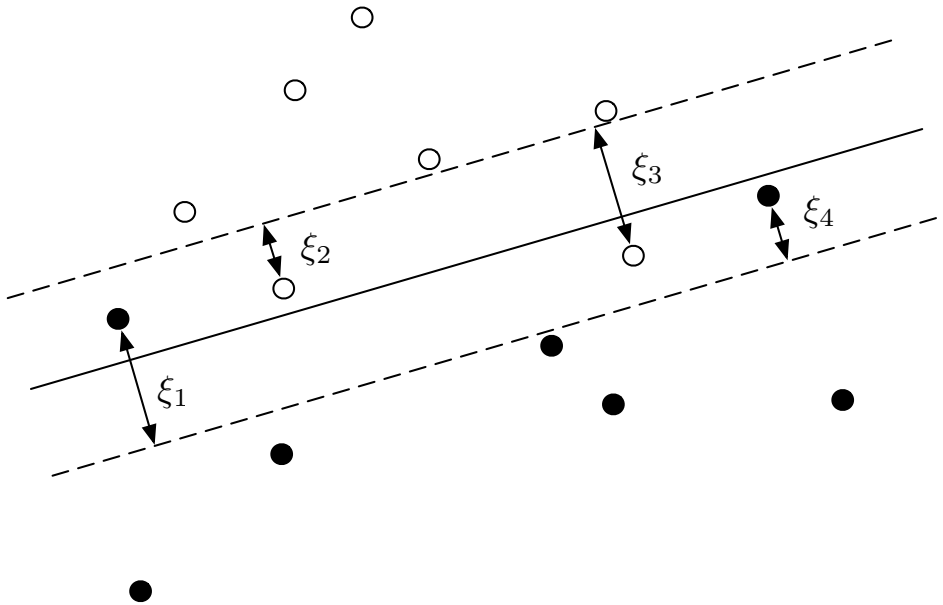
- We just presented a method to minimize the structural risk in linear machines by explicitly limiting the VC dimension of the hyperplane.
- The previous classification problem is not real because the data is linearly separable. On it, the empirical risk is just zero.
- In real cases, the data is not linearly separable, and an empirical risc must be defined and minimized.
- The SVM idea consists of defining an empirical risk and minimizing it together with the maximization of the margin in a functional.

- Consider a nonseparable set of data $\mathcal{D}$ and slack variables ξ_n as

$$y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) > 1 - \xi_n$$

with $\xi_n \geq 0$.

- They are zero for data outside the margin. Also, in the figure, $\xi_1, \xi_3 > 1$ and $\xi_2, \xi_4 < 1$.



- A way to minimize the empirical risk consists of minimizing the sum of slack variables:

$$R_{emp}(\mathbf{w}) = \sum_{n=1}^{N} \xi_n$$

$$\text{subject to } y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) > 1 - \xi_n$$

- The Support Vector Machine idea consists of minimizing the previous empirical risk plus the structural risk through margin maximization, this is:

$$\text{minimize } L_p(\mathbf{w}, \xi_n) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{subject to } y_n\left(\mathbf{w}^\top \mathbf{x}_n + b\right) > 1 - \xi_n$$

$$\xi_n \geq 0$$

- C is a free *tradeoff* parameter.
- Subindex p stands for *primal*. We'll have a *dual* later.
- The name of Support Vector Machine will become apparent later.

Conclusion

The following concepts have been reviewed in this lesson:

- The VC dimension h.
- The VC theorem that introduces the trade off between empirical risk and structural risk through h.
- The structural Risk Minimization principle that leads to the SVM.
- The relationship between the margin maximization criterion and the minimization of the VC dimension.
- The SVM functional to be optimized.